Uso de una Discretización de la Transformada de Fourier para Identificación de Individuos por Voz

José Francisco Rico Andrade y José Antonio Camarena Ibarrola

***Resumen*—La identificación de un individuo por su voz puede hacerse comparando el espectrogama de una palabra pronunciada por este, con los espectrogramas de las elocuciones de la misma palabra pronunciada por los individuos identificables por el sistema. Para que el uso de espectrogramas produzca buenos resultados en el problema de identificación de individuos es importante una determinación precisa de los contenidos de frecuencia de la señal de voz, la cual es una señal no estacionaria. En este artículo, se propone el uso de una Discretización de la Transformada de Fourier Continua de reciente aparición para tal efecto. En los experimentos realizados, se obtuvieron mejores tazas de reconocimiento que los obtenidos al usar la Transformada de Fourier Discreta.**

***Temas claves*—Muestreo de Señales, Polinomios de Hermite, Trasformación de Similitud, Transformada de Fourier, Densidades de Potencia Espectral, Doblado Dinámico en Tiempo, Reales Positivos, Falsos Positivos, Gráficas de Características Operativas de Recepción.**

I. Introducción

Los procesos de identificación de individuos por voz, han empleado tradicionalmente técnicas basados en análisis espectral de señales para la obtención de parámetros característicos. [1]. Algunas, como Densidades de Potencia Espectral (PSD - Power Spectral Densities) y Coeficientes Cepstrales en las Frecuencias de Mel (MFCC - Mel Frequency Cepstral Coefficients), dependiendo de su implementación [7], se han vuelto populares debido a algunos de sus resultados [3] [4] [6]. Dichas técnicas dependen del análisis de Fourier, para el proceso de extracción de parámetros característicos [5] [2], y se han realizado de manera conveniente sobre procesadores digitales. Sin embargo, en la práctica, la Transformada de Fourier Discreta (DFT - Discrete Fourier Transform), asume que una señal es periódica, tomando dicha señal completa como un periodo de una forma de onda periódica [10]. Sin embargo, las señales de voz no son periódicas, particularmente cuando se trata de sonidos fricativos (consonantes que generan señales ruidosas) [8]. Por su parte, la Transformada de Fourier Continua (CFT - Continuos Fourier Transform) no tiene esta restricción, sin embargo, no es posible obtener una función continua en el tiempo, que describa las señales de voz. No obstante, recientemente se ha definido una Discretización de la CFT en [11] [12], y se ha propuesto un algoritmo para su utilización en la discriminación de sonidos vocalizados de fricativos [9].

En este artículo se propone una adaptación al algoritmo propuesto en [9] para el cálculo de CFT, que permita obtener espectros de frecuencia equiparables a los obtenido por DFT, de manera que sea posible comparar los resultados de ambos algoritmos, en un esquema de identificación de individuos por voz texto dependiente, utilizando Doblado Dinámico en Tiempo (DTW - Dynamic Time Warping) [13] [14] [15], como medida de alineamiento, y análisis de las gráficas de Características Operativas de Recepción (ROC - Receiver Operating Characteristics) [19], como técnica, para organizar los clasificadores y visualizar su desempeño.

II. Cálculo de CFT.

En [9] se propone un algoritmo para el análisis de señales de voz que a groso modo consta de los siguientes pasos:

1. Encontrar los coeficientes de un polinomio trigonométrico que se ajusten a la forma de onda de la señal de voz.

2. Formar un vector con las raíces de un polinomio de Hermite de grado elevado.

3. Construir una matriz con el kernel de Fourier acorde a [11] [12].

4. Evaluar el polinomio trigonométrico encontrado en el paso 1, en el vector con las raíces del polinomio de hermite del paso 2.

5. Calcular la discretización de la CFT multiplicando la matriz del paso 3 con el vector obtenido en el paso 4.

Sin embargo, para realizar una comparación de espectros obtenidos mediante DFT y este algoritmo, existen algunos inconvenientes, destacando: La complejidad de la búsqueda de los coeficientes del polinomio trigonométrico, que mejor se ajusten a la forma de onda de la señal de voz. Así como el hecho de que con dicho algoritmo solo es posible la obtención de un espectro de frecuencias en gamas de frecuencia relativamente pequeñas (aunque de gran resolución), que varían con la cantidad de raíces del polinomio de Hermite de grado *P* utilizadas. Aunando a esto, se encuentra la complejidad para encontrar las raíces de polinomios de Hermite de grado elevado.

Por ello, en este artículo se propone una adecuación de dicho algoritmo, como se detalla en la **Sección II.A**.

J. A. Camarena labora en la Universidad Michoacana de Sán Nicolás de Hidalgo, Ciudad Universitara, Edif. Ω2 , CP 58004, Morelia, Mich., México. (e-mail: camarena@umich.mx).

J. F. Rico labora en la Universidad Michoacana de Sán Nicolás de Hidalgo, Ciudad Universitara, Edif. Ω1 , CP 58004, Morelia, Mich., México. (e-mail: jfrico@correo.fie.umich.mx).

A. Algoritmo Propuesto.

1. Formar un vector xh_p, con las raíces del polinomio de Hermite de grado *P* = 3000

2. Considerando un vector d_{xh}, de longitud *P*-1, con las distancias entre los valores adyacentes del vector xh_p, donde $d_{xh}[i]=\left|xh_p[i+1]-xh_p[i]\right|$, obtener un vector rd_{xh} con las razones geométricas, o cocientes, de cada uno de los valores $rd_{xh}[i]$ entre la distancia mínima en d_{xh} (distancia de la raíz positiva de menor valor absoluto a la raíz negativa de menor valor absoluto); haciendo $rd_{xh}[i]=d_{xh}[i]/\min\{d_{xh}\}$. Así, se utiliza un factor arbitrario $\alpha_{xh}=1.0219$, para discriminar razones de distancia en rd_{xh}, y considerar únicamente aquellas que cumplan con la restricción $rd_{xh}<\alpha_{xh}$, para formar así un vector xh_c, con los C valores centrales de xh_p, considerados equidistantes; **Fig. 1**.

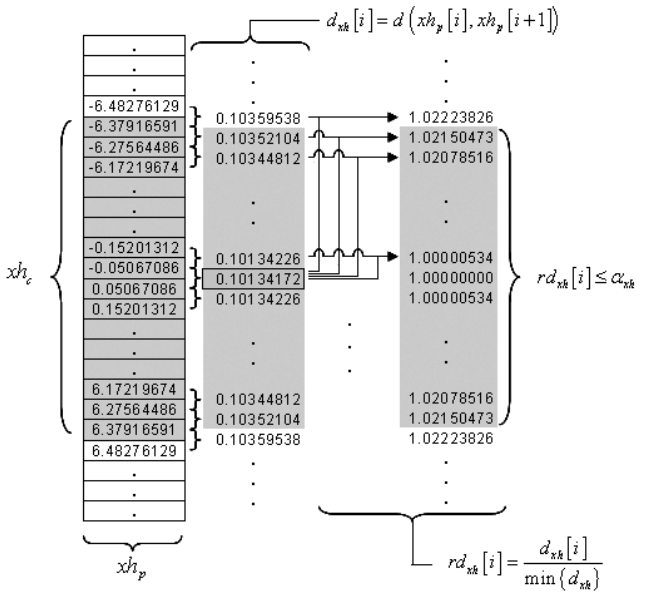

Fig. 1 Obtención de xh_c, a partir de xh_p

3. Obtener un vector xh_{ck}, con los valores de xh_c, agregando en cada extremo, $k=99$ pseudo-raíces (ver **Fig 2**), de manera que sea posible representar un rango de frecuencias análogo al obtenido por FFT, para una misma señal. Donde cada pseudo-raíz se calcula con la media de las distancias entre todas las parejas de valores adyacentes en xh_c.

4. Siendo r_w un marco de *N*=256 muestras de longitud, de la señal de voz. Obtener un sub-muestreo xh_{sm} de $S=techo\left((C+2k)-s\right)$, de los valores en xh_{ck}, considerando solo uno de cada *s* de sus elementos (con la intención de disminuir la necesidad de aproximar 980 valores a partir de las 256 muestras de cada marco de la señal de voz), como se detalla en la **Fig. 3**.

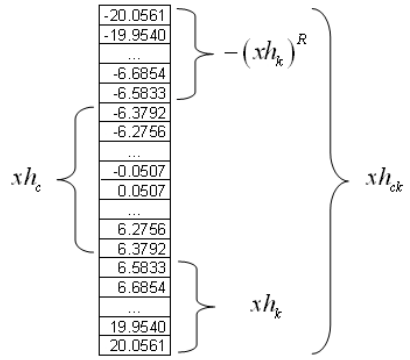

Fig 2 Descomposición de xh_{ck}**.**

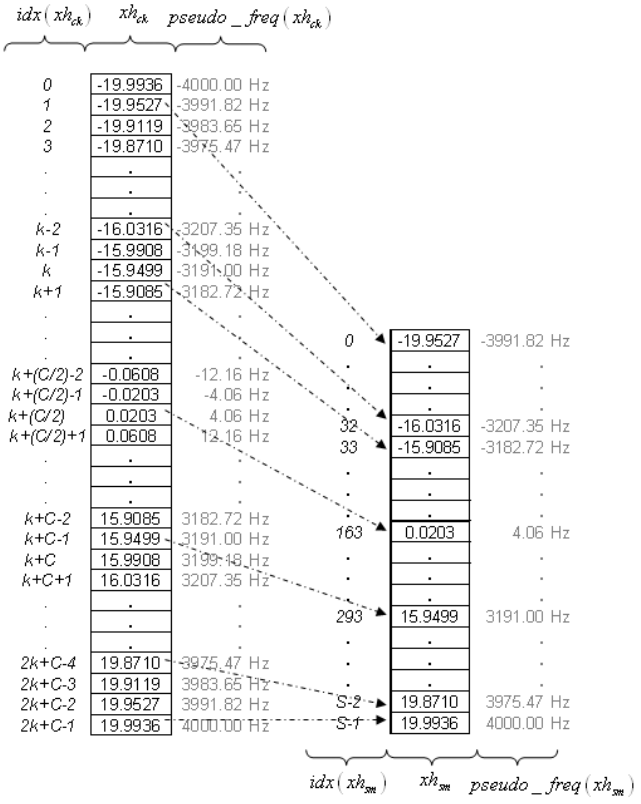

Fig. 3. Vector xh_{sm} **, para** *P* = 3000**,** *C* = 782**,** *k* = 99 **y** *s*=3

5. Re-escalar la secuencia de pseudo-raíces $xh_{sm}[0], xh_{sm}[1],\ldots, xh_{sm}[S\text{-}1]$, del vector xh_{sm}, en valores de 0 a *N*-1, llamando xh_{sc} al vector resultante. El proceso anterior puede considerarse como se describe en los pasos **5.1** y **5.2**

5.1 Obtener xh_{smd} resultante de aplicar un desplazamiento de xh_{sm}, haciendo $xh_{smd}=xh_{sm}+xh_{sm}[0]$ de modo que $xh_{smd}[0]=0$ **Fig 4**.

5.2 Obtener xh_{sc} resultante de aplicar a xh_{smd}, una transformación de similitud haciendo $xh_{sc}=xh_{smd}\bullet\left(N-1/xh_{smd}[S-1]\right)$, de manera que $xh_{sc}[S-1]=N$. **Fig 4**.

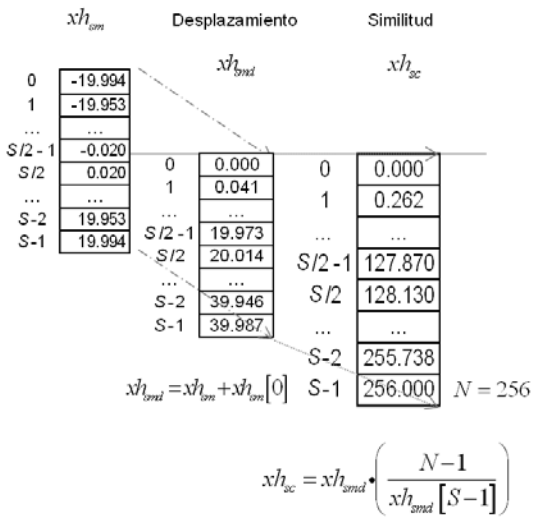

Fig 4 Re-escalamiento xh_{sc} de xh_{sm}.

6. Construir el vector r_{sc} re-muestreando r_w, en los valores de xh_{sc}, mediante interpolación lineal.

7. Construir la matriz *F*, utilizando la ecuación (1):

$$F_{m,n} = \frac{\pi}{\sqrt{2p}} \sqrt[4]{\frac{4p+3-xh_{sm}^2[n]}{4p+3-xh_{sm}^2[m]}} \left[\cos(xh_{sm}[m]\, xh_{sm}[n]) + j \sin(xh_{sm}[m]\, xh_{sm}[n]) \right] \quad (1)$$

donde *F* de dimensiones *m* , *n*, es la representación finita del kernel de Fourier, *j* es la unidad imaginaria, y xh_{sm}, el vector con el sub-muestreo de las pseudo-raíces del polinomio de Hermite, utilizando *P* = 3000.

8. Calcular $g = F \cdot r_{sc}$, donde *g*, será la discretización de la CFT de r_{sc}.

III. Doblado Dinámico en Tiempo.

Alinear dos secuencias en tiempo, *T* y *P*, de longitud *N* y *M* respectivamente, para las cuales: $T = [t_0, t_1, ..., t_i, ..., t_N]$, y $P = [p_0, p_1, ..., p_j, ..., p_N]$; equivale a encontrar una función de doblado $j = W(i)$, que mapea índices *i* y *j* de manera que se obtenga el registro entre las secuencias *T* y *P*. Dicha función *W* se restringe tradicionalmente a las condiciones de los límites *W*(0) = 0 y *W*(*N*) = *M*, y puede sujetarse a diversas condiciones locales, con pesos diferentes para cada una de ellas [13].

Generalmente [16] [17] [18], para alinear estas dos secuencias, DTW construye una matriz *D* de dimensión *N* , *M*, representando las distancias de la mejor ruta parcial posible, considerando las condiciones locales elegidas. Generalmente se consideran aquellas para las cuales, si la función de doblado óptimo, lleva al punto $D_{i,j}$, la ruta óptima debe pasar por $D_{j-1,i}$, $D_{j-1,i-1}$, $D_{j,i-1}$, con un peso para cada restricción, como se muestra en la **Fig. 5**

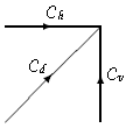

Fig. 5 Restricciones locales simétricas de primer orden.

De esta manera, una vez seleccionadas las restricciones locales y sus pesos, puede crearse la matriz *D*, en base a la inicialización (2) y (3), así como la ecuación recursiva (4).

$$D_{i,0} = \sum_{k=0}^{i} d(t_k, p_0) \quad (2)$$

$$D_{0,j} = \sum_{k=0}^{j} d(t_0, p_k) \quad (3)$$

$$D_{i,j} = mín \begin{cases} D_{i-1,j} + c_v * d(t_i, p_j) \\ D_{i,j-1} + c_h * d(t_i, p_j) \\ D_{i-1,j-1} + c_d * d(t_i, p_j) \end{cases} \quad (4)$$

donde $d(t_i, p_j)$, representa la distancia entre los marcos t_i y p_j, para $1 \le i \le N$ y $1 \le j \le M$; c_x representa el peso de la restricción *x*. De este modo, el alineamiento que resulta de la minima distancia entre las dos secuencias, tiene el valor $D_{N,M}$.

IV. Análisis ROC.

Las curvas ROC presentan una gráfica bidimensional en la cual la razón de reales positivos (aciertos), se grafica sobre el eje de las ordenadas y la razón de falsos positivos (errores), sobre el eje de las abscisas.

En la mayoría de los clasificadores, existe un parámetro que se puede ajustar (umbral de decisión), de modo que se incrementen/decrementen las razones de reales positivos y falsos positivos, respectivamente. Cada clasificador, para un valor específico de este parámetro, produce un par de dichas razones, que corresponde con un único punto en el espacio ROC. Así, una variación del parámetro de clasificación (umbral de decisión), puede utilizarse para graficar una curva ROC como en la **Fig. 6**. Donde informalmente, resulta preferible, aquel umbral, que mejor se aproxime al clasificador perfecto (punto de la curva a menor distancia de (0, 1)). En adición, se considera deseables aquellos clasificadores que se encuentren más a la izquierda (sección conservadora) del espacio ROC, que aquellos a la derecha (sección liberal), puesto que solo realizan clasificaciones cuando poseen suficiente información para hacerlo. [19].

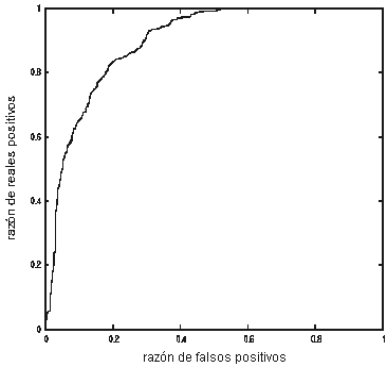

Fig. 6: Curva ROC para un clasificador paramétrico

Un método común, para comparar clasificadores es utilizar también, el área bajo la curva ROC, abreviado AUC (Area Under the Curve) [19].

V. EXPERIMENTOS.

Se utilizan la implementación propuesta de la discretización de la CFT y el algoritmo de Cooley-Tukey para el cálculo rápido de la DFT [5], como técnicas de extracción de parámetros característicos en la clasificación de individuos por voz bajo un esquema texto dependiente, considerando una población de 21 locutores y un diccionario de 34 palabras (dígitos del cero al nueve y alfabeto griego del alfa al omega), considerando 3 elocuciones (denominadas como: a, b y c, para su diferenciación), por cada individuo, para cada una de ellas. Cada elocución fue registrada en un archivo WAV PCM sin compresión, muestreado a 8000Hz, cuantizado a 16bits, monaural. Para las comparativas, se consideran como entrada, conjuntos de 3 x 21 archivos (una misma palabra, por cada locutor), del total de los 2142 archivos del corpus de elocuciones. La señal de cada uno de ellos es procesada en bloques de 256 muestras, transformada al dominio de la frecuencia, calculado su nivel de presión sonora como se describe en [21], y agrupada en las primeras 18 bandas bark (unidades de la escala psicoacústica de frecuencia perceptual del mismo nombre [20]), para obtener su espectrograma (se considera un traslape entre marcos de de *2/3* de *N*). En base a dichos espectrogramas se calculan matrices de confusión [19], utilizando DTW como medida de distancia entre elocuciones. Con lo anterior, se calculan tanto positivos (falsos y verdaderos), como negativos y con ellos las curvas ROC correspondientes a cada implementación de la Transformada de Fourier para su comparación.

A. Utilizando DTW para el cálculo de distancias entre elocuciones y llenado de las Matrices de Confusión..

Se implementó DTW utilizando las ecuaciones (2), (3), y (4), para las restricciones locales mostradas en la **Fig. 5** con los pesos: $c_h = 1$, $c_d = 2$ y $c_v = 1$. Normalizando la distancia DTW obtenida, mediante $D_{N,M} / (N + M)$.

Con dichos parámetros se construye una matriz D_E de distancias, de tamaño $N_E \times N_E$, por cada palabra registrada en el diccionario que conforma el corpus de elocuciones; siendo N_E, el número total de elocuciones a probar. De modo que $D_E[i, j] = DTW(E_i, E_j)$; de esta manera, cada fila *i* en D_E, representa el vector de distancias DTW obtenidas considerando E_i como elocución de prueba, y E_j , para $j = 0,1,...,N_E - 1$ como elocuciones de referencia.

Las **Figs 8**, **9** y **10**, muestran las matrices de distancias $D_{Epsilon\_DFT}$, $D_{Epsilon\_CFT\_s=1}$ y $D_{Epsilon\_CFT\_s=4}$, respectivamente, considerando Negro para distancia 0, y blanco para la mayor distancia; obtenidas para la palabra Epsilon (considerando las 3 instancias a, b y c registrada de la misma palabra por cada locutor), caracterizadas por DFT, CFT s=1 y CFT s=4.

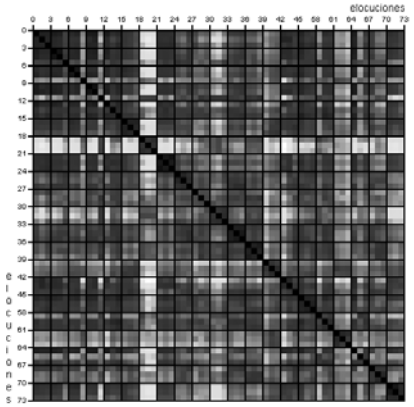

Fig. 8. Elocuciones Caracterizadas por DFT.

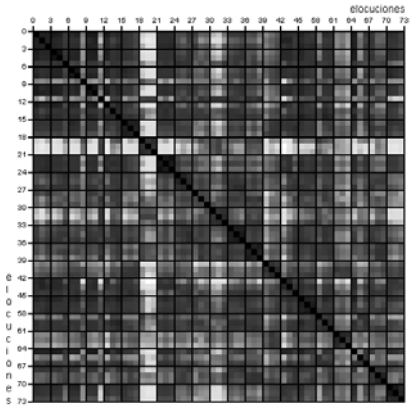

Fig. 9. Elocuciones Caracterizadas por CFT s=1.

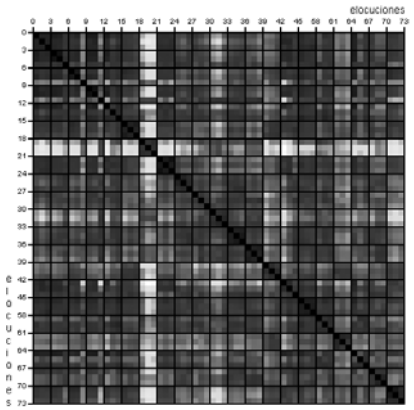

Fig. 10. Elocuciones Caracterizadas por DFT s=4.

La **Tabla I** muestra el ordenamiento de las elocuciones, empleado para la obtención de las matrices de distancias mostradas en las **Figs 8**, **9** y **10**.

Así, estableciendo umbrales de distancia entre elocuciones caracterizadas, se obtienen las razones de reales positivos y falsos positivos para utilizar en el análisis ROC. Para lo que utilizando el corpus de elocuciones descrito en la sección 5, se obtuvieron por cada técnica de análisis espectral, un espectrograma por cada archivo y en base a estos, se calculó el área bajo la curva ROC, correspondiente a la identificación de los 21 individuos considerados en el corpus de elocuciones.

TABLA I
LOCUTORES Y ORDENAMIENTO EMPLEADO DE ELOCUCIONES PARA MATRICES DE CONFUSIÓN POR DISTANCIAS

Locutor	Numero de elocución	Elocución
Aaron	01	Épsilon – a
	02	Épsilon – b
	03	Épsilon – c
Alfredo	04	Épsilon – a
	05	Épsilon – b
	06	Épsilon – c
Alma	07	Épsilon – a
	08	Épsilon – b
	09	Épsilon – c
Rafael	10	Épsilon – a
	11	Épsilon – b
	12	Épsilon – c
Daniel	13	Épsilon – a
	14	Épsilon – b
	15	Épsilon – c
Eréndira	16	Épsilon – a
	17	Épsilon – b
	18	Épsilon – c
Ernesto	19	Épsilon – a
	20	Épsilon – b
	21	Épsilon – c
Florentino	22	Épsilon – a
	23	Épsilon – b
	24	Épsilon – c
Gustavo	25	Épsilon – a
	26	Épsilon – b
	27	Épsilon – c
Iván	28	Épsilon – a
	29	Épsilon – b
	30	Épsilon – c
Jesús	31	Épsilon – a
	32	Épsilon – b
	33	Épsilon – c
Jorge	34	Épsilon – a
	35	Épsilon – b
	36	Épsilon – c
Julio	37	Épsilon – a
	38	Épsilon – b
	39	Épsilon – c
Kristel	40	Épsilon – a
	41	Épsilon – b
	42	Épsilon – c
Merari	43	Épsilon – a
	44	Épsilon – b
	45	Épsilon – c
Mario	46	Épsilon – a
	47	Épsilon – b
	48	Épsilon – c
Mauritz	49	Épsilon – a
	50	Épsilon – b
	51	Épsilon – c
Nayeli	52	Épsilon – a
	53	Épsilon – b
	54	Épsilon – c
René	55	Épsilon – a
	56	Épsilon – b
	57	Épsilon – c
Alejandro	58	Épsilon – a
	59	Épsilon – b
	60	Épsilon – c
Vivianne	61	Épsilon – a
	62	Épsilon – b
	63	Épsilon – c

VI. RESULTADOS.

Como se señaló en la **sección II.A** la longitud de un marco r_{sc} así como el vector g de coeficientes de frecuencia (obtenidos por discretización de CFT propuesta), es inversamente proporcional al parámetro s de sub-muestreo empleado, como se detalla en la **Tabla II**.

Así, a partir de $s = 4$ la cantidad de muestras consideradas por discretización de CFT es menor a las 256, consideradas con FFT. Sin embargo $s = 4$, aproxima de mejor manera la cantidad de muestras analizadas por ambos algoritmos.

TABLA II
COEFICIENTES DE FRECUENCIA PARA DIFERENTES VALORES DE *S*

Transformada	*s*	*Muestras*
FFT	N/A	256
CFT	1	980
CFT	2	490
CFT	3	327
CFT	4	245
CFT	5	196
…	…	…

Este fenómeno altera los resultados del reconocimiento de individuos, provocando variaciones irregulares e inconsistentes, en las áreas bajo la curva ROC conforme se incrementa el valor de s.

En base a lo anterior, se obtuvo el promedio de las áreas bajo las curvas ROC, de cada elocución del corpus; tanto para FFT, como para discretizaciones de CFT con $1 \le s \le 25$ (**Fig. 11**), donde se puede apreciar como el espacio en el que varían las áreas bajo la curva ROC para discretización de CFT, tiende a incrementarse a medida que lo hace también el parámetro *s*; aún cuando en promedio, el área bajo la curva ROC, para discretización de CFT, tienda a mantenerse por encima del área obtenida con FFT (excepto para $s = 25$).

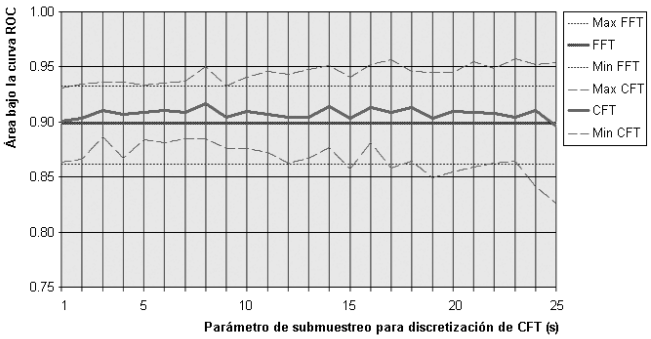

Fig. 11. Promedio de áreas bajo las curvas ROC para FFT y discretizaciones de CFT.

Considerando las 34 palabras (100%) de que está compuesto el corpus de elocuciones, se encontró que para s=4, el algoritmo propuesto para discretización de CFT, permite para el 94.1% de ellas, obtener resultados superiores a las caracterizaciones obtenidas por FFT. Siendo así 4, el valor de *s* que mejor caracteriza, una mayor cantidad de palabras en el corpus, como se muestra en la **Fig. 12**.

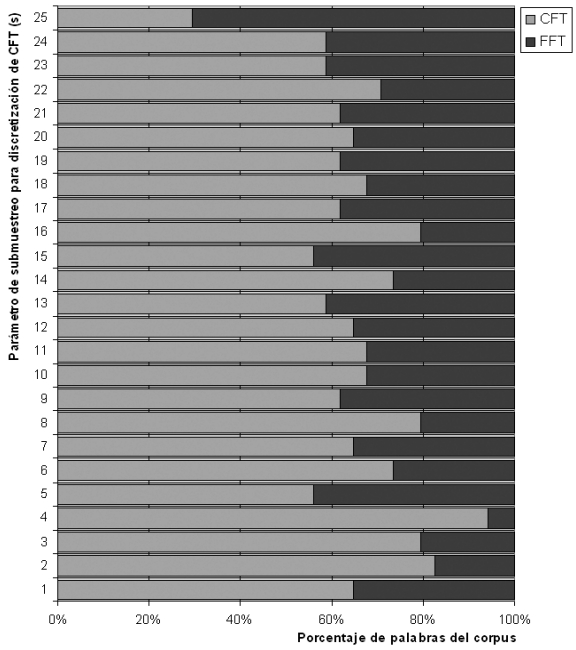

Fig. 12. Porcentajes de palabras con mejor área bajo la curva para FFT y discretizaciones de CFT

La **tabla 3**, detalla los porcentajes de palabras que mejor clasificación obtuvieron, tanto para FFT como para CFT con *s=4*.

TABLA III
PALABRAS MEJOR CLASIFICADAS POR FFT Y DISCRETIZACIÓN DE CFT PARA S=4

CFT con *s=4*	FFT
Cero, Uno, Tres, Cuatro, Cinco, Seis, Siete, Ocho, Nueve, Alfa, Beta, Gamma, Delta, Epsilon, Dseta, Eta, Kappa, Lambda, Mi, Ni, Xi, Omicron, Pi, Ro, Sigma, Tao, Upsilon, Fi, Ji, Psi, Omega	Dos, Teta

En la **Fig. 13** se muestra gráficamente los resultados obtenidos en las áreas bajo las curvas ROC, correspondientes a cada palabra analizada, tanto con discretización de CFT para s= 4, como con FFT.

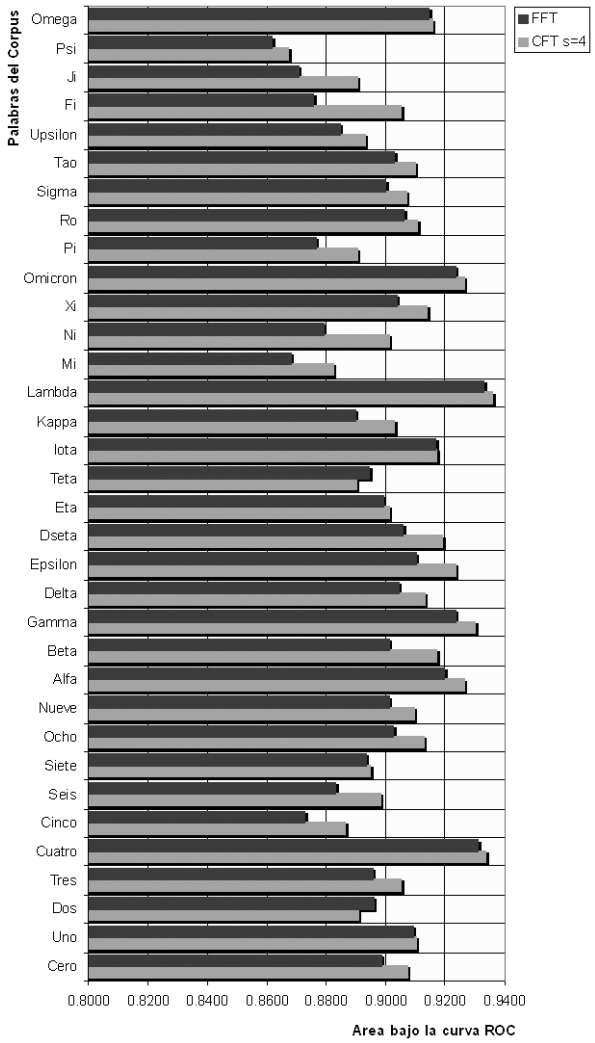

Fig. 13. AUCs ROC para las palabras del corpus empleando FFT y discretización de CFT con s=4.

A partir de los datos de la Fig. 13, se calculó el promedio del área bajo las curvas ROC del total de palabras del corpus, tanto para FFT, como para CFT con *s*=4. Los resultados de dichos promedios se muestran en la **Fig. 14**, observando así, que las áreas obtenidas con CFT para *s*=4, son superiores tanto para 32 de las 34 palabras del corpus, como para el promedio de áreas bajo la curva, y el rango en que éstas varían.

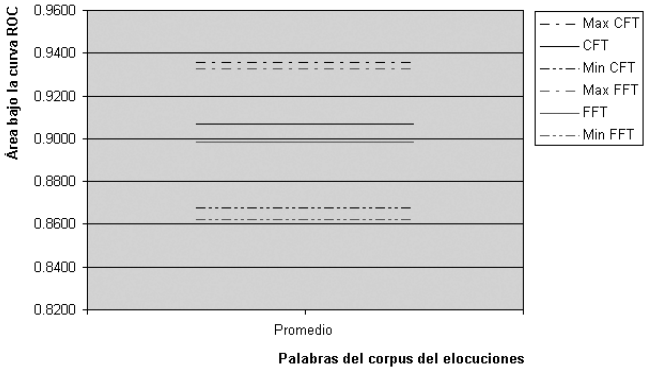

Fig. 14. Promedio de AUCs ROC para las palabras del corpus empleando FFT y CFT con *s=4*.

Ahora bien, si se analizan a detalle las curvas ROC de las dos palabras para las que el área obtenida por CFT no fue superior a la obtenida por FFT, puede observarse que CFT con s = 4, presenta una superioridad en la región conservadora del espacio ROC, y decrementando su elevación solo en la sección liberal de dicho espacio, con respecto de la curva obtenida con FFT (**Figs 15** y **16**).

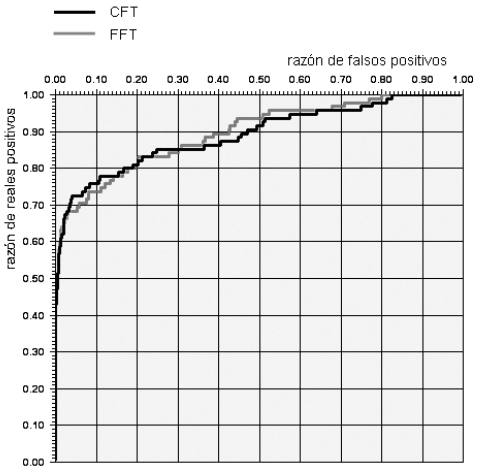

Fig. 15. Curvas ROC para palabra "Dos"

VII. CONCLUSIONES.

La discretización de CFT implementada, a través de los diferentes submuestreos, para uno de cada $1 \le s \le 25$ valores de xh_{ck}, mostró un comportamiento inestable para las diferentes palabras del corpus, en el análisis ROC, correspondiente a la identificación de individuos por voz. Sin embargo, fue posible identificar, valores de dicho parámetro, para los que discretización de CFT, presentó mejoras considerables, con respecto de las áreas bajo la curva ROC, obtenidas mediante DFT. En adición, se analizaron a detalle los resultados correspondientes al utilizar el parámetro *s*=4,

con el cual, discretización de CFT superó las AUCs obtenidas con FFT, en un 94.1% del corpus. Donde del 5.9% restante se observó que es posible elegir un umbral de clasificación que arroje mejores resultados de los que se obtendría con FFT.

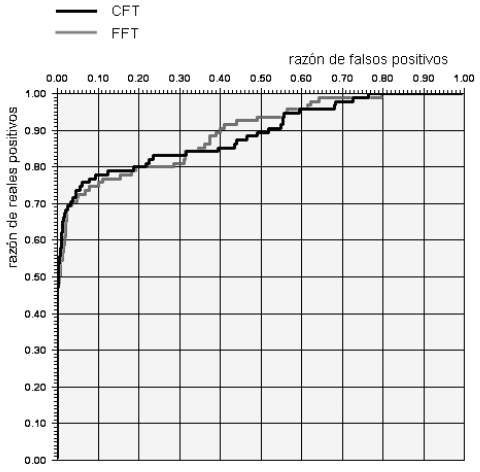

Fig 16. Curvas ROC para palabra "Teta"

Sin embargo, aunque se presentaron algunas mejoras en el análisis ROC, para discretización de CFT con respecto de FFT, al emplearse como técnica de caracterización de señales de voz; resulta evidente que la complejidad del algoritmo propuesto O(n^2), es muy superior al orden O(nLogn) de la FFT. De donde se desprende la necesidad de realizar estudios posteriores, que profundicen en las propiedades de la discretización de CFT, para estar en posibilidades de aprovecharla de una mejor manera.

VIII. Referencias.

[1] Sadaoki Furui, "Survey of the State of the Art in Human Language Technology", NTT Human Interface Laboratories, Tokyo, Japan, 1996, http://cslu.cse.ogi.edu/HLTsurvey/ch1node9.html, 05 de Octubre de 2008.

[2] ETSI. Speech processing, transmission and quality aspects (stq) - distributed speech recognition; - front-end feature extraction algorithm. ETSI ES 201 108-v1.1.3-2003-09, 2003. URL: http://www.etsi.org/services_products/freestandard/home.htm, Consulta: Agosto 2006.

[3] Ganchev, T., Fakotakis, N., y Kokkinakis, G. Comparative evaluation of various mfcc implementations on the speaker verification task. 10th International Conference on Speech and Computer (SPECOM 2005), Vol. 1:pp. 191-194, 2005.

[4] Gupta, H., Hautamäki, V., Kinnunen, T., y Fränti, P. Field evaluation of text-dependent speaker recognition in an access control application. Proceedings of the 10th International Conference Speech and Computer (SPECOM'2005), 2005.

[5] Press, W. H., Teukolsky, S. A., Vetterling, W. T., y Flannery, B. P. Numerical Recipes in C: The Art of Scienti□c Computing. Cambridge University Press, 2a edón., 1992. ISBN 0-521-43108-5.

[6] Schalk, H., Reininger, H., y Euler, S. "A system for text dependent speaker verification - □field trial evaluation and simulation results. Proceedings of EUROSPEECH-2001, págs. pp. 783-786, 2001.

[7] Zheng, F., Zhang, G., y Song, Z. "Comparison of diferent implementations of MFCC". J. Computer Science & Technology, Vol. 16(6):pp. 582-589, 2001.

[8] Sankaranarayanan, A. A text-dependent approach to speaker identification. TechOnline, Audio Design Line, 2002. URL: http://www.audiodesignline.com/showArticle.jhtml?articleID=192200467 , Consulta: Diciembre 2006.

[9] Camarena-Ibarrola, A. y Chávez, E. Using a new discretization of the fourier transform to discriminate voiced from unvoiced speech. 7o Encuentro Internacional de Ciencias de la Computacion, Paper # 74 San Luis Potosi, Septiembre de 2006, 2006. URL: http://lc.fie.umich.mx/~camarena/ENC2006_74.pdf, Consulta: Septiembre 2006.

[10] Proakis, J. y Manolakis, D. G. Digital signal processing: principles, algorithms and applications. Macmillan Publishing Company, 2ª ed. 1996. ISBN 0133942899.

[11] Campos, R. G. y Z., L. J. A discretization of the continuos fourier transform. Nuovo Cimento, Vol. 106B. pp. 703-711, 1992.

[12] Campos, R. G. A quadrature formula for the hankel transform. Numerical Algorithms, Vol. 9 (No. 3-4). pp. 343-354, 1995.

[13] Sakoe, H. y Chiba., S. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoustics, Speech and Signal Processing - ASSP, Vol. 26:pp. 43 - 49, 1978.

[14] Rabiner, L. R., Rosenberg, A. E., y Levinson, S. E. Considerations in dynamic time warping algorithms for discrete word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 26(No. 6):pp. 575 - 582, 1978.

[15] Itakura, F. Minimum prediction residual principle applied to speech recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 23:pp. 52 - 72, 1987.

[16] Robinson, J. Pseudo-local alignment of continuous time series. Duke University, 2004. URL: http://files.fangengine.org/~josh/pdf/josh230.pdf Consulta: 27 de Agosto de 2006

[17] Tomasi, G., van den Bergand, F., y Andersson, C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. JOURNAL OF CHEMOMETRICS, Vol. 18:pp. 231 - 241, 2004.

[18] Camarena-Ibarrola, J. A. y Chávez, E. Identifying music by performances using an entropy based audio-fingerprint. Mexican International Conference on Artificial Intelligence (MICAI), 2006.

[19] Fawcett, T. Roc graphs: Notes and practical considerations for researchers. Machine Learning, 2004. URL citeseer.ist.psu.edu/fawcett04roc.html

[20] Zwicker, E. y Fastl, H. Psychoacoustics: Facts and Models. Information Sciences. Springer, 2ª ed., 1999. ISBN 3-540-65063-6.

[21] Bosi, M. y Goldberg, R. E. Introduction to Digital Audio Coding and Standards. Engineering and Computer Science. Kluwer Academic Publishers, 1ª ed., 2002. ISBN 1-4020-7357-7.

IX. Biografías

J. Antonio Camarena I. nació en Morelia, Michoacán México el 11 de Julio de 1964. Se graduó como Ingeniero Electricista en la UMSNH, luego como Maestro en Ciencias Computacionales en el Instituto Tecnológico de Toluca y posteriormente obtuvo el grado de Doctor en Ciencias en Ingeniería Eléctrica Opción Sistemas Computacionales de la UMSNH.

Su experiencia profesional incluye Administración de Sistemas y de Base de Datos en el Instituto Federal Electoral y Profesor-Investigador de la UMSNH miembro de la división de estudios de postgrado de la Facultad de Ingeniería Eléctrica. Sus áreas de interés incluyen, entre otras el reconocimiento de patrones y la caracterización de señales de audio.

J. Francisco Rico A. nació en Morelia, Michoacán México, el 18 de marzo de 1980. Se graduó del Instituto Tecnológico de Morelia, como Ingeniero en Sistemas Computacionales, y posteriormente como Maestro en Ciencias en Ciencias de la Computación.

Su experiencia profesional incluye la compañía Dealer de México empresa de grupo FAME, el Sistema Michoacano de Radio y Televisión. Ha trabajado como docente en el Instituto Tecnológico de Morelia, Universidad Vasco de Quiroga y la Facultad de Eléctrica de la Universidad Michoacana de San Nicolás de Hidalgo. Sus áreas de interés incluyen, entre otras, diseño de sistemas web, Bases de Datos, Redes de Computadoras, Teleproceso, Procesamiento Digital de Señales.